МИР 18.09.2025 | Мустафа Сулейман

Эпоха «разумных» машин близко

ИИ еще не обрел сознания, но уже кажется таким. Именно эта иллюзия может стать опасным поворотом в развитии технологий.

Миссия всей моей жизни: создать безопасный и полезный искусственный интеллект (ИИ), который сделает мир лучше. Но сейчас у меня вызывает растущую озабоченность перспектива, что люди так сильно поверят в искусственные интеллекты как сознательные сущности, что начнут бороться за «права ИИ» и даже за их гражданство. Это стало бы опасным поворотом в развитии технологий. Его необходимо предотвратить. Мы должны создавать ИИ-модели для людей, а не чтобы они были людьми.

В этом смысле споры о том, может ли ИИ реально обладать сознанием, являются отвлекающим маневром. В ближайшей перспективе важна лишь иллюзия сознания. Мы уже близки к системам, которые я называю «кажущийся сознательным ИИ» (сокращенно SCAI): они будут достаточно убедительно имитировать сознание.

SCAI сможет свободно пользоваться естественным языком и убедительно демонстрировать личность, вызывающую эмоциональный отклик. У него будет долгая и точная память, формирующая целостное представление о себе, и он будет пользоваться этой способностью для претензий на субъектный опыт (ссылаясь на предыдущие контакты и воспоминания). Сложные функции вознаграждения внутри этих ИИ-моделей будут стимулировать внутреннюю мотивацию, а продвинутая постановка целей и планирование будут усиливать наше ощущение, что ИИ обладает реальной агентностью.

Мы должны задуматься о последствиях, утвердив нормы, которые запретят

Все эти возможности либо уже есть, либо скоро появятся. Мы должны признать, что такие системы могут возникнуть уже очень скоро, и задуматься о последствиях, утвердив нормы, которые запретят создание иллюзорного сознания у ИИ.

создание иллюзорного сознания у ИИ

Уже сейчас многие люди считают взаимодействие с ИИ богатым и реальным жизненным опытом, приносящим чувство удовлетворения. Усиливаются тревоги по поводу «ИИ-психоза», формирования чувства привязанности и влияния на психическое здоровье. Сообщается даже, что некоторые люди воспринимают ИИ как проявление Бога. А ученые, специализирующиеся на вопросах сознания, рассказывали мне, что завалены запросами людей, желающих знать, обладает ли их ИИ сознанием, и можно ли в него влюбиться.

Тот факт, что SCAI технически возможен, мало что говорит нам о возможности обладания этими системами сознанием. Как отмечает нейробиолог Анил Сет, симуляция грозы в компьютере не означает, что там реально идет дождь. Когда инженеры создают внешние маркеры сознания, они не воссоздают реальное сознание. Но мы должны признать, что на практике появятся люди, которые смогут создавать SCAI-системы, утверждающие, будто они реально обладают сознанием. И что еще важнее, найдутся люди, которые им поверят, воспринимая маркеры сознания за *действительное* сознание.

Даже если это воспринимаемое сознание не будет реальным (а по этому вопросу будут вестись бесконечные споры), социальный эффект, конечно, будет реальным. Сознание тесно связано с нашим чувством идентичности и пониманием моральных и юридических прав внутри общества. Если кто-нибудь создаст SCAI, и эта ИИ-система убедит людей в своей способности страдать или в том, что у нее есть право не быть выключенной, тогда появятся люди, которые станут добиваться их защиты. В мире, который уже погружен в поляризующие споры о правах и идентичности, возникнет новая трещина – между теми, кто за и против прав ИИ.

Опровергнуть утверждения, что ИИ страдает, будет непросто из-за ограниченности имеющихся научных данных. Некоторые ученые уже исследуют идею «благополучия моделей», и доказывают, что наш «долг – распространять соображения морали на сущности, у которых имеется не пренебрежимо малый шанс... наличия сознания».

Применять такой принцип преждевременно и опасно. Это усугубит заблуждения излишне восприимчивых людей, злоупотребляя их психологической уязвимостью. Кроме того, усложнится

деятельность правозащитников, поскольку будет создана новая, огромная категория обладателей прав. Именно поэтому появление SCAI нужно обязательно предотвратить. Наше внимание следует сосредоточить на защите благополучия и прав людей, а также животных и окружающей среды.

Пока что мы не готовы к грядущим новшествам. Опираясь на растущий корпус различных исследований взаимодействия людей с ИИ, нам срочно нужно установить четкие нормы и принципы. Один из таких принципов: ИИ-компании не должны поощрять веру в то, что их ИИ обладает сознанием.

ИИ-отрасль (да и вся техноотрасль в целом) нуждается в надежных принципах разработок и в наилучших методах противодействия такого рода иллюзиям. Например, специально задуманные сбои позволят разрушить иллюзию, мягко напомнив пользователям об ограниченности ИИ-систем и их истинной природе. Подобные протоколы должны специально формулироваться и разрабатываться (возможно, по требованию закона).

Реально усиливающий людей ИИ должен быть максимально полезным и минимально имитирующим сознание

В компании Microsoft AI мы проявляем инициативу, пытаясь понять, как могла бы выглядеть ответственная «ИИ-личность», и какие ограничения у нее должны быть. Эта работа имеет фундаментальное значение, поскольку для устранения рисков SCAI нужна позитивная версия ИИ-компаньонов, дополняющих нашу жизнь без вреда для здоровья.

Мы должны стремиться создавать ИИ-модели, которые будут побуждать людей к восстановлению связей между собой в реальном мире, а не к бегству в параллельную реальность. А если взаимодействие с ИИ носит длительный характер, модели должны всегда представляться только как ИИ, а не как люди-фейки. Реально усиливающий людей ИИ должен быть максимально полезным и минимально имитирующим сознание.

Перспективой появления SCAI надо заняться немедленно. Во многих отношениях это будет момент, когда ИИ вдруг станет резко полезным: он сможет управлять инструментами, запоминать все детали нашей жизни и так далее. Но нельзя игнорировать риски этих способностей. Мы все знаем увлекающихся людей, готовых погрузиться «в кроличью нору». Это будет вредно для их здоровья – и для здоровья общества.

Если ИИ будет создаваться с явной целью быть похожим на людей, он будет терять свой истинный потенциал – инструмента, который усиливает людей.

(c) Project Syndicate 2025

Мустафа Сулейман

Мустафа Сулейман (Mustafa Suleyman) – генеральный директор Microsoft AI и автор книги *The Coming Wave: Technology, Power, and the Twenty-First Century's Greatest Dilemma* (Crown, 2023). Ранее он был сооснователем компаний Inflection AI и DeepMind.